BIOMETRICS 44, 417–431
June 1988

Effects of Treatment-Induced Mortality and Tumor-Induced Mortality on Tests for Carcinogenicity in Small Samples

A. John Bailer and Christopher J. Portier

Division of Biometry and Risk Assessment,
National Institute of Environmental Health Sciences, P.O. Box 12233,
Research Triangle Park, North Carolina 27709, U.S.A.

SUMMARY

Statistical tests of carcinogenicity are shown to have varying degrees of robustness to the effects of mortality. Mortality induced by two different mechanisms is studied—mortality due to the tumor of interest, and mortality due to treatment independent of the tumor. The two most commonly used tests, the life-table test and the Cochran–Armitage linear trend test, are seen to be highly sensitive to increases in treatment lethality using small-sample simulations. Increases in tumor lethality are seen to affect the performance of commonly used prevalence tests such as logistic regression. A simple survival-adjusted quantal response test appears to be the most robust of all the procedures considered.

1. Introduction

The problem of comparing treated groups in the presence of potential confounders is common in statistics. Various statistical procedures have been used to address this problem, including stratified analyses where strata are formed by grouping levels of the confounding variables and regression procedures where the confounding variables are controlled in a continuous fashion. This general problem will be discussed in the context of animal studies where carcinogenesis as a function of treatment level is the outcome of interest and survival time is the confounding factor.

The basic study design for determining carcinogenicity is the lifetime exposure rodent study. In a typical carcinogenicity experiment, rodents are exposed to some compound from the time of weaning (approximately 6–8 weeks old) until they die or are sacrificed at the end of the study (usually 2 years). The data obtained from each animal include the presence or absence of a tumor(s) and the age at death. The cause of death will be assumed to be unknown. The question of interest in these studies is whether a significant dose-response relationship exists between the administered compound and the presence of some tumor or tumors. One major difficulty in such studies is that the time of tumor onset is not observable. Various assumptions concerning the effect of tumor onset on animal survival are used to circumvent this difficulty. Survival times confound the question of carcinogenicity when the different treatment groups experience different patterns of censoring with respect to tumor onset.

Two major factors that influence survival are treatment-induced mortality (or treatment lethality) and tumor-induced mortality (or tumor lethality). Treatment lethality refers to the potential of high doses of a compound to shorten survival. For this discussion, treatment lethality refers to nontumor mortality. In carcinogenicity studies, substances are given at

Key words: Carcinogenicity; Hoel–Walburg test; Logistic regression test; Quantal response; Survival model; Trend tests.

high doses that can be toxic, and thus might lead to the animal's early death before a tumor is observed. Tumor lethality refers to the influence of tumor presence on survival. A highly lethal tumor would cause death shortly after its onset. An incidental tumor is a tumor whose presence does not affect survival. It is readily apparent that both tumor lethality and treatment lethality will affect survival times.

Statistical procedures applied in this context make assumptions regarding tumor lethality to allow inference on tumor onset times. Standard life-table analyses are appropriate for tumors that are considered to be lethal (Mantel, 1966; Tarone, 1975). Logistic regression techniques (Dinse and Lagakos, 1983) or the Mantel–Haenszel-type procedure of Hoel and Walburg (1972) are appropriate for incidental tumors. This paper addresses the properties of these tests under varying degrees of tumor lethality and treatment lethality when cause of death is unattributable. In Section 2, the statistical tests to be considered are described in detail. Section 3 describes a simulation experiment using a simple stochastic model of tumor onset for visualizing the progression from a tumor-free state to a death state with or without the tumor. In Section 4, the results of the simulation study are presented, and our results are summarized in the last section.

2. Tests of Carcinogenesis

Consider a carcinogenicity experiment with K treated groups and a control group where animals in the ith group receive dose z_i, $i = 0, 1, \ldots, K$. Suppose strata have been formed over S time intervals, and let n_{is} denote the number at risk in the sth stratum for the ith group, and let d_{is} denote the number of animals with the response of interest in this same group–stratum combination. A dot in a subscript will be used to denote summation over that subscript; hence, $n_{i.} = \Sigma_s n_{is}$, $d_{i.} = \Sigma_s d_{is}$, $n_{.s} = \Sigma_i n_{is}$, and $d_{.s} = \Sigma_i d_{is}$. Let ε_{is} denote the expected number of animals that have the response of interest in the sth stratum of the ith group. Using a multinomial model, we find that ε_{is} can be estimated by $E_{is} = n_{is}(d_{.s}/n_{.s})$. Most of the commonly used tests for carcinogenicity can be formulated as special cases of a generic test statistic Z_{G}, where

$$Z_{\mathrm{G}} = \frac{T_{\mathrm{G}}}{V_{\mathrm{G}}}$$

with

$$T_{\mathrm{G}} = \sum_{i=0}^{k} z_i(d_{i.} - E_{i.})$$

and

$$V_{\mathrm{G}}^2 = \sum_{s=1}^{S} V_s^2 = \sum_{s=1}^{S} \left(\frac{d_{.s}}{n_{.s}}\right)\left(\frac{n_{.s} - d_{.s}}{n_{.s}}\right)\left(\frac{n_{.s}}{n_{.s} - 1}\right) \sum_{i=0}^{k} n_{is}(z_i - \bar{z})^2,$$

$$\bar{z} = (n_{..})^{-1} \sum_{i=0}^{K} n_{i.} z_i.$$

This type of generic test statistic formulation has also been used elsewhere (Gart et al., 1987, Chap. 5; Bailer, Institute of Statistics Mimeo Series #1815T, University of North Carolina, Chapel Hill, 1986). The different tests are oriented toward different null hypotheses depending on the assumptions used in deriving the test. However, in all cases, under the null hypothesis, Z_{G} is asymptotically distributed as a standard normal variate. The distinction between most of the various test statistics is in the formation of the stratum and the definition of the number at risk.

2.1 *Onset Test*

Throughout this paper, we will use the term *tumor onset* to mean the presence of the first histopathologically detectable tumor of the type being studied. We will not be concerned with multiple tumors. Assuming that tumors are irreversible allows us to treat tumor onset in the same manner as we treat a death in an ordinary survival analysis. If tumor onset were observable, the life-table test derived by Tarone (1975) would be applicable to testing the hypothesis of equal tumor incidence rates in all groups. The random generation of tumor onset time is an intermediate step in the computer simulation of such carcinogenicity experiments. Therefore, this test is applicable in the framework of simulation studies. This "onset test" corrects for survival and requires no assumptions concerning tumor lethality; thus, it provides a useful standard with which to compare the other tests. In the tumor onset test, strata are defined by each tumor onset time. Animals that die prior to getting a tumor are considered as censored observations. The animals given dose z_i that are alive and tumor-free just prior to the sth tumor onset time are denoted by n_{is}, and d_{is} represents the number of animals who subsequently get the tumor at the sth tumor onset time.

2.2 *Quantal Response*

The remaining tests can be divided into two broad categories: binomial-based tests that use only the quantal response (though two of the quantal response procedures discussed below attempt to address potential survival differences) and tests that correct for treatment-induced survival differences.

The Cochran–Armitage linear trend test (Armitage, 1955) considers the data from the dose groups to be collapsed over the entire study period into one stratum ($S = 1$). All tests of this type define d_{i1} to be the number of animals in group i that are found to have the tumor. The Cochran–Armitage linear trend test uses n_{i1} equal to the number of animals placed on study and Z_G can be calculated accordingly. An implicit assumption in the use of this test is that all animals are at equal risk of getting the tumor over the duration of the study. However, because tumors sometimes have long latency periods and because some treatments decrease survival, animals may die earlier in some treatment groups and thus be at decreased risk of tumor onset. One method of correcting for this problem would be to modify the value of n_{i1} to reflect less-than-whole-animal contributions for decreased survival. One way of doing this would be to define the number at risk for these tests as

$$n_{i1}^* = \sum_{j=1}^{n_{i1}} \omega_{ij}$$

and then test hypotheses of increasing trends in the modified proportions, r_i, where this modified proportion is of the form

$$r_i = \frac{d_{i1}}{n_{i1}^*}.$$

The weights, ω_{ij}, are all equal to 1 for the Cochran–Armitage trend test.

Gart et al. (1979) suggest a procedure that defines the weights as $\omega_{ij} = 1$ if the age at death for the jth animal in the ith group, t_{ij}, exceeds the time of the first death with tumor present and $\omega_{ij} = 0$ if not. This test will be referred to as the *truncated trend test*.

As a second adjustment of this type, we will define the weights as $\omega_{ij} = 1$ if the jth animal in the ith group dies with the tumor present and $\omega_{ij} = (t_{ij}/t_{\max})^3$ if not, where $t_{\max}$ is the maximum survival time. This weighting scheme results from the observation that many tumors seem to appear at the rate of a third- to fifth-order polynomial in time (see e.g., Portier, Hedges, and Hoel, 1986). This test will be referred to as the *Poly*-3 *trend test*. Further motivation for this test will be discussed later.

2.3 *Survival-Adjusted Tests*

Life-table test The life-table test is similar to the onset test except that strata are now formed for each death time with a tumor, and the animals that are sacrificed at the conclusion of the study form their own stratum. Animals that die without a tumor are treated as censored observations. This is analogous to the combined analysis of fatal tumors and terminal sacrifice tumors suggested by Peto et al. (1980). In terms of the generic test statistic, n_{is} is the number of animals in group i that are alive just prior to the sth death and d_{is} is the number of animals dying with the tumor present at the sth death time. This is a test of treatment-related differences in the hazard of death with the tumor.

Prevalence tests Hoel and Walburg (1972) proposed to test the hypothesis of equal tumor prevalence in the dosed groups within strata that are chosen external to the observed survival times. The National Toxicology Program uses five strata when applying this procedure (Haseman, 1984), where the strata are the time intervals (expressed in weeks): 0–52, 53–78, 79–92, 93–Terminal sacrifice, and Terminal sacrifice (often 104 weeks). For this procedure, n_{is} is the number of animals dying in the interval formed by stratum s in group i and d_{is} is the number of these animals with the tumor present.

Logistic regression (Dinse and Lagakos, 1983) can be used to model tumor prevalence as a function of dose and survival time. Unlike the Hoel–Walburg procedure, logistic regression adjusts for survival times in a continuous manner. In what follows, we use a logistic model with a linear time effect and a linear treatment effect, and the hypothesis of equal tumor prevalence is tested with a score test as employed by Dinse (1985). Note that the generic test statistic Z_G has been shown by Birch (1965) to be asymptotically equivalent to a logistic regression score test under a model with a linear treatment effect and a time parameter for each stratum. Hitchcock (1966) demonstrated near equivalence in finite samples.

3. A Simulation Experiment

It is convenient to use a three-state stochastic model to describe the results of a carcinogenicity experiment when cause of death is unobtainable. This model is illustrated in Figure 1. The transition rates in this model are described by the hazard functions $\lambda_i(t)$, $\beta_i(t)$, and $\gamma_i(t, \omega)$. Let E_1 be the random variable that represents the time from initial exposure to the occurrence of the first event which is either a tumor onset or a tumor-free death. Let E_2 represent the time until tumor-bearing death. Let δ be an indicator of tumor presence ($\delta = 1$) or absence ($\delta = 0$). The hazard functions in Figure 1 can be defined as

$$\lambda_i(t) = \lim_{\Delta\downarrow 0}(\Delta)^{-1}\Pr\{t \leqslant E_1 < t + \Delta,\ \delta = 1 \mid E_1 \geqslant t,\ z_i\};$$

$$\beta_i(t) = \lim_{\Delta\downarrow 0}(\Delta)^{-1}\Pr\{t \leqslant E_1 < t + \Delta,\ \delta = 0 \mid E_1 \geqslant t,\ z_i\};$$

$$\gamma_i(t, \omega) = \lim_{\Delta\downarrow 0}(\Delta)^{-1}\Pr\{t \leqslant E_2 < t + \Delta \mid E_1 = \omega \leqslant t,\ \delta = 1,\ E_2 \geqslant t,\ z_i\}.$$

Our interest concerns whether increases in administered dose are related to increases in tumor incidence. As discussed by McKnight and Crowley (1984), the null hypothesis in carcinogenicity studies should be expressed in terms of tumor incidence rates because tests in terms of other rates can be biased when test assumptions are violated. In the context of

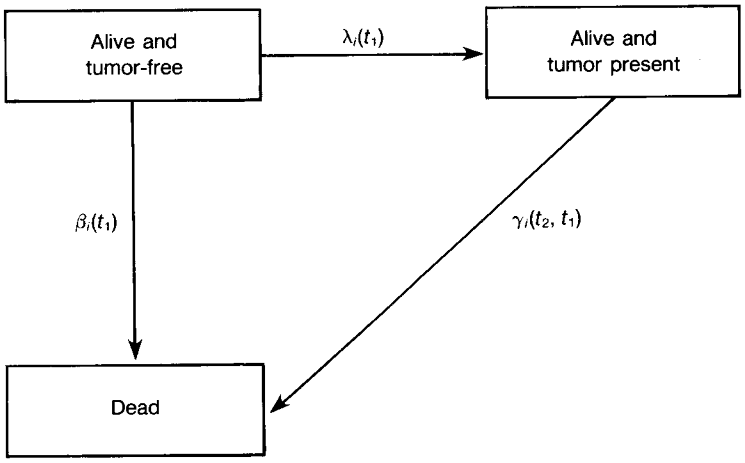

Figure 1. Three-state stochastic model for carcinogenicity experiments when cause of death is unobtainable.

the model above, the null hypothesis test of interest then would be H_0: $\lambda_0(t) = \lambda_1(t) = \cdots = \lambda_K(t)$. The incidence test addresses this hypothesis directly because it is based on the time of tumor onset. The remaining tests address this hypothesis in special cases as outlined below.

The hypothesis tested by the Cochran–Armitage trend test is whether the probability of developing the tumor in group i before study termination (TS), π_i, is the same in every group. This probability can be expressed in terms of the hazard functions defined above as

$$\pi_i = \int_0^{TS} \lambda_i(u) \exp\left\{-\int_0^u [\lambda_i(x) + \beta_i(x)]\, dx\right\} du. \tag{1}$$

Under the assumption that $\lambda_0(t) = \cdots = \lambda_K(t)$, the only time we are assured that $\pi_0 = \pi_1 = \cdots = \pi_K$ is when $\beta_0(t) = \beta_1(t) = \cdots = \beta_K(t)$. In other cases, it is possible to have $\pi_j \neq \pi_i$ $(i \neq j)$ even when $\lambda_0(t) = \lambda_1(t) = \cdots = \lambda_K(t)$. Similarly, it is possible to have $\pi_j = \pi_i$ $(i \neq j)$ when the λ's differ.

It is shown in the Appendix that the modified proportion r_i as defined in Section 2.2 approximates

$$1 - \exp\left[-\int_0^{TS} \lambda_i(s)\, ds\right],$$

where ω_{ij} is chosen to be

$$\omega_{ij} = \frac{\int_0^{t_{ij}} \lambda_i(s)\, ds}{\int_0^{TS} \lambda_i(s)\, ds}$$

for animals that die at time t_{ij} without the tumor. When $\lambda_i(s)$ is of the order of t^2, the Poly-3 test satisfies this condition, and the test should be fairly robust to varying degrees of treatment lethality. The Appendix gives arguments as to why this approximation should work well. The simulations that follow address this point directly for small samples.

The hypothesis tested in the life-table test is that there is no difference in the event-specific hazard functions for death with tumor in the $K + 1$ groups (Tarone, 1975). The hazard is "event-specific" because differences in the time until death with tumor present are of interest. Death without the tumor present is treated as a censoring mechanism. The hazard of death at time s in group i and tumor onset before time s will be denoted by $h_i(s)$. The hazard function, $h_i(s)$, can be written in the terms of the stochastic model as

$$h_i(s) = \frac{\int_0^s H_i(u, s)\gamma_i(s, u)\,du}{\int_0^s H_i(u, s)\,du + \exp\{-\int_0^s [\lambda_i(x) + \beta_i(x)]\,dx\}}, \tag{2}$$

where

$$H_i(u, s) = \lambda_i(u)\exp\left\{-\int_0^u [\lambda_i(x) + \beta_i(x)]\,dx\right\}\exp\left[-\int_u^s \gamma_i(w, u)\,dw\right].$$

The term $H_i(u, s)$ represents the probability that an animal with tumor onset at time u survives until time s. When the assumption of instantly lethal tumors is true [i.e., $\Pr(D = t \mid T = s) = 1$ for $t = s$ and 0 for $t > s$], the expression for the hazard function reduces to $h_i(s) = \lambda_i(s)$.

The life-table test rejects the null hypothesis of no survival differences in the various dose groups when decreased survival with tumor is associated with increasing dose levels. If the hazard of death given the tumor is present is some function of the hazard of death given the tumor is absent [i.e., $\gamma_i(t, u) = f(\beta_i(t), u)$] then decreased survival with tumor in the higher-dose groups can occur when $\beta_0(t) \leqslant \beta_1(t) \leqslant \cdots \leqslant \beta_K(t)$ even though $\lambda_0(t) = \lambda_1(t) = \cdots = \lambda_K(t)$. Therefore, treatment lethality effects can cause greater than expected Type I errors and inflated power estimates.

The prevalence tests compare the probabilities of tumor presence given death within a particular stratum or at a certain time. The null hypothesis for these procedures can be stated as H_0: $\kappa_0(s) = \cdots = \kappa_K(s)$ where $\kappa_i(s) = \Pr(\text{tumor present} \mid \text{animal dies at time } s)$. This probability will equal the prevalence, Pr(tumor present | animal alive at time s), if the tumor is nonlethal. In terms of the stochastic model given above, this probability can be expressed as

$$\kappa_j(s) = \frac{A_i(s)}{A_i(s) + \beta_i(s)\exp\{-\int_0^s [\lambda_i(x) + \beta_i(x)]\,dx\}}, \tag{3}$$

where $A_i(s)$, the density for death with tumor at time s, is given by

$$A_i(s) = \int_0^s \exp\left\{-\int_0^u \lambda_i(x) + \beta_i(x)]\,dx\right\}\lambda_i(u)\exp\left[-\int_u^s \gamma_i(w, u)\,dw\right]\gamma_i(s, u)\,du.$$

Under the incidental tumor assumption, $\gamma_i(w, u) = \beta_i(w)$, this probability reduces to

$$\kappa_j^*(s) = \frac{A_i^*(s)}{A_i^*(s) + \exp[-\int_0^s \lambda_i(x)\,dx]},$$

where

$$A_i^*(s) = \int_0^s \lambda_i(u)\exp\left[-\int_0^u \lambda_i(x)\,dx\right]du.$$

Hence, the prevalence tests are functions only of the tumor incidence rate when the incidental tumor assumption is valid.

From looking at these null hypotheses, we get some indication of the possibility for incorrect Type I error rates using these tests. The simulation study that follows was used to quantify this problem. A major concern in the simulation of survival data is determining a reasonable form for the hazard functions described above and finding logical values for the parameters of these functions. Portier et al. (1986) provide hazard functions that adequately fit a large historical database of untreated animals. The models suggested by their analyses are used in this study. For the control group tumor incidence rate, we use a two-parameter Weibull function of the form

$$\lambda_0(t) = \eta_1 \eta_2 t^{\eta_2 - 1}.$$

For the control group hazard of death given the tumor is absent, a modified Weibull hazard function provided an adequate fit. This model is given by

$$\beta_0(t) = \alpha_1 + \alpha_2 \alpha_3 t^{\alpha_3 - 1}.$$

For the purposes of this simulation, we will treat $\beta(t)$ as not varying with tumor type. The hazard functions for the other groups will be modelled in a proportional hazards framework where

$$\lambda_i(t) = (1 + \eta_0 z_i)\lambda_0(t) \quad \text{and} \quad \beta_i(t) = (1 + \alpha_0 z_i)\beta_0(t).$$

The values for α_i and η_i ($i \geqslant 1$) that were used in the following simulations can be found in Portier et al. (1986). A representative sample of three sex–species–tumor site combinations will be discussed in detail in the following sections. These three combinations were selected because they illustrate a range of background tumor rates from approximately 1% to 19%.

The hazard of death for a tumor-bearing animal was modelled by assuming that this hazard was equal to the hazard of death for a tumor-free animal plus some treatment-independent continuous function of the time since tumor onset; i.e., $\gamma_i(t, \omega) = \beta_i(t) + f(t - \omega)$. The portion of the cumulative hazard attributed to f can be written as

$$\int_{\omega}^{t} f(s - \omega)\, ds = \phi_0 \alpha_2 (t - \omega)(t_{\max})^{\alpha_3 - 1}.$$

This cumulative hazard implies that the lifetime hazard of death for an animal with early tumor onset is approximately $(1 + \phi_0)$ times the hazard of death for a tumor-free animal. Thus, by varying the value for ϕ_0, we can range from incidental tumors ($\phi_0 = 0$) to highly lethal tumors ($\phi_0 \gg 0$).

Two experimental designs were considered in the simulations: a four-dose design with 50 animals per group and doses of 0, .25, .50, and 1; and a three-dose design with 50 animals per group and with doses of 0, .50, and 1. The results from these different designs were very similar; hence, only the results from the four-dose design are reported. In this simulation study, four onset factors (η_0), three treatment lethality factors (α_0), and three tumor lethality factors (ϕ_0) were considered. Hence, 36 onset by treatment lethality by tumor lethality combinations were studied for each of 20 sex–species–tumor site combinations yielding a total of 720 unique simulation conditions. Each simulation condition was replicated 1,300 times, which leads to an approximate 95% confidence interval of [.038, .062] for a true binomial probability of rejection of .05.

The three treatment lethality effects considered were $\alpha_0 = 0$, $\alpha_0 = 1$, and $\alpha_0 = 4$. Treatment lethality effects of $\alpha_0 \geqslant 1$ were found in approximately 30% of the gavage experiments considered in unpublished work by Bailer and Portier. Treatment lethality effects of $\alpha_0 \geqslant 5$ were observed in over 10% of the gavage experiments in this same set of experiments. Bailer and Portier also considered feeding studies. In 11% of those studies, a treatment lethality of $\alpha_0 \geqslant 1$ was observed.

4. Results

4.1 *Type* I *Error Rates*

The estimated Type I error rates of the test statistics for a subset of the models considered in this simulation study are presented in Table 1. The three sex–species–tumor site

Table 1

Type I error of carcinogenicity tests for varying levels of treatment lethality and tumor lethality using a nominal .05 level

Sex–species; tumor rate	Tumor site	Treatment lethality[a] $1 + \alpha_0$	Tumor lethality[b] $1 + \phi_0$	Test statistic[c]						
				Z_o	Z_t	Z_p	Z_{tt}	Z_{lt}	Z_{hw}	Z_{log}
Female rats 1.2%	Lung	1	1	4.3[d]	4.2	4.2	4.2	4.3	4.3	4.1
			2	3.7	3.7	3.8	3.7	3.9	4.0	3.6
			10	4.3	4.3	4.3	4.3	4.5	4.8	4.7
		2	1	4.1	3.8	3.9	5.7	6.7	5.5	5.6
			2	4.0	3.9	3.9	5.5	6.4	5.0	5.1
			10	3.5	3.2	3.5	3.9	4.4	3.2	3.7
		5	1	5.2	4.3	9.8	11.2	14.3	6.5	10.6
			2	3.7	3.2	8.0	8.1	11.8	4.4	7.4
			10	3.5	3.3	9.0	5.2	6.5	2.0	4.5
Male rats 4.6%	Liver	1	1	5.1	5.0	4.9	5.2	5.5	5.1	5.2
			2	4.6	4.9	4.7	4.5	4.3	4.8	4.4
			10	4.8	4.8	4.7	4.9	4.3	4.6	4.8
		2	1	6.5	4.5	5.7	5.8	8.7	5.5	6.1
			2	5.7	4.6	5.0	5.8	8.8	5.6	5.9
			10	4.5	3.3	3.9	3.8	6.0	2.4	3.1
		5	1	5.8	1.5	3.9	6.1	18.2	4.9	6.6
			2	5.8	1.7	3.8	5.2	17.7	5.0	6.0
			10	5.8	1.8	4.2	4.2	12.2	2.5	3.6
Female rats 19.1%	Leuk./ lymphoma	1	1	5.3	5.7	5.2	5.5	5.5	5.4	5.4
			2	4.0	4.2	3.8	4.2	4.2	4.2	4.0
			10	6.0	5.8	5.9	5.9	6.1	5.3	4.8
		2	1	5.1	3.3	4.9	4.3	11.2	4.8	5.1
			2	5.1	3.7	4.9	4.0	10.2	4.2	4.2
			10	4.6	3.5	4.4	3.5	7.8	1.8	2.7
		5	1	5.7	1.2	5.0	2.5	43.2	5.0	6.2
			2	5.5	1.2	5.5	1.9	38.3	3.2	3.5
			10	5.5	1.9	5.6	2.2	20.5	.5	.8

[a] Treatment lethality: $1 + \alpha_0 = 1$ (no treatment lethality); = 2 (low lethality); = 5 (high treatment lethality).

[b] Tumor lethality: $1 + \phi_0 = 1$ (incidental tumor); = 2 (low tumor lethality); = 10 (high tumor lethality).

[c] Test statistics: Z_o = Onset test; Z_t = Cochran–Armitage trend test; Z_p = Poly-3 test; Z_{tt} = Truncated trend test; Z_{lt} = Life-table test; Z_{hw} = Hoel–Walburg test; Z_{log} = Logistic regression score test.

[d] Entries are given in percentages.

combinations presented in Table 1 provide a representative sample of the background tumor rates we considered.

The Type I error rate of the incidence test, Z_o, is unaffected by changes in treatment lethality or by changes in tumor lethality.

The Cochran–Armitage trend test, Z_t, the Poly-3 trend test, Z_p, and the truncated trend test, Z_{tt}, have smaller Type I error rates with increasing treatment lethality for almost all conditions considered. This result confirms the statements in the previous section where the π_i and Ω_i (Appendix) are seen to be a function of both $\lambda_i(x)$ and $\beta_i(x)$. In fact, π_i is a decreasing function of $\beta_i(x)$. Therefore, when $\lambda_0(t) = \lambda_1(t) = \cdots = \lambda_K(t)$ and $\beta_0(t) \leqslant \beta_1(t) \leqslant \cdots \leqslant \beta_K(t)$, we see that $\pi_0 \geqslant \cdots \geqslant \pi_K$, and the Cochran–Armitage trend test will be less likely to reject than nominally specified. Tumor lethality seems to have no effect on the significance levels of these tests. This is an expected result because π_i and Ω_i (Appendix) are not functions of γ_i. Generally, both of the modified trend tests (Z_p and Z_{tt}) show superior robustness to treatment lethality over that of the Cochran–Armitage trend test. The modified trend procedures tend to have larger Type I error rates than the true significance level for the smallest background tumor rate, and the effect of treatment lethality on the level of the quantal response trend tests appears to increase as the background tumor rate increases. The results for the Poly-3 test support the arguments given in the Appendix that this test should be robust with respect to varying degrees of treatment lethality. In addition, we observed that the test statistic seems to have the correct distributional form for the small samples considered.

The conservative response of Z_t under increasing treatment lethality conditions may be explained as follows. Since treatment lethality may serve as a censoring agent, animals in the higher-dose groups may be less likely to live long enough for tumor onset than animals in the control group. Thus, the resulting estimate of the probability of tumor response in the high-dose groups may be artificially low relative to the estimate in the control group. Under the null hypothesis of no linear trend in the probability of tumor onset, this would lead to Z_t rejecting less frequently than nominally specified. This fact becomes obvious when one considers equation (1).

As seen in Table 1, the observed Type I error rate was inflated in the truncated trend test, Z_{tt}, for rare tumors and high treatment lethality. A possible explanation for this phenomenon is that the sample sizes used in calculating the probabilities of tumor onset for Z_{tt} may be dramatically reduced in situations of rare tumors. With high treatment lethality and rare tumors, the number at risk may be very small in the higher-dose groups, which could lead to inflated estimates of the probability of tumor onset in those groups. These inflated estimates in the higher-dose groups could cause Z_{tt} to reject more frequently than nominally specified under the null hypothesis. Gart, Chu, and Tarone (1979) suggested including in the number at risk only those animals that died after some predetermined age for rare tumors. This adjustment may improve the operating characteristics of this test for rare tumors and high treatment lethality.

When considering the operating characteristics of the Poly-3 trend test, Z_p, one must discuss the factor $(t_{ij}/t_{\max})^3$ when the true shape, η_2, of the onset distribution differs from $\eta_2 = 3$. If $\eta_2 < 3$ (e.g., female rat lung tumors), the factor $(t_{ij}/t_{\max})^3$ will be smaller than it would be if η_2 were used as an exponent. Thus, the number at risk would be smaller using the exponent equal to 3, which in turn implies that the estimate of the probability of tumor onset would be larger than it would be if the true onset shape parameter were used. In the case where $\eta_2 < 3$ and some treatment lethality is present, it might be predicted that Z_p would reject more frequently than nominally specified under the null hypothesis. By analogy, one might predict that Z_p would reject less frequently than nominally specified for $\eta_2 > 3$ and for some treatment lethality. The predictions from this hypothetical discussion are confirmed by the results presented in Table 1.

Table 2

Power of carcinogenicity tests for varying levels of treatment lethality and tumor lethality when the treatment induces a twofold tumorigenic effect in the high-dose group

Sex–species; Tumor rate	Tumor site	Treatment lethality[a] $1 + \alpha_0$	Tumor lethality[b] $1 + \phi_0$	Test statistic[c] Z_o	Z_t	Z_p	Z_{tt}	Z_{lt}	Z_{hw}	Z_{log}
Female rats 1.2%	Lung	1	1	13.5[d]	13.4	13.5	13.6	13.1	13.2	13.5
			2	12.2	12.2	12.0	12.4	11.9	11.8	12.0
			10	11.2	11.1	11.2	11.1	11.1	10.6	12.2
		2	1	10.6	10.2	11.2	12.9	15.2	11.8	12.8
			2	12.2	12.0	12.8	13.8	16.1	11.8	12.7
			10	10.0	9.8	10.4	10.3	11.3	6.8	8.7
		5	1	11.0	9.4	16.3	16.7	27.0	10.3	14.8
			2	12.9	11.2	17.5	16.8	25.8	10.4	14.1
			10	12.8	11.2	18.3	13.3	17.3	4.0	7.7
Male rats 4.6%	Liver	1	1	23.8	23.8	23.9	23.6	23.6	24.0	23.2
			2	20.5	20.6	20.2	20.8	20.5	20.3	20.7
			10	22.8	22.4	22.5	22.8	22.7	22.1	22.5
		2	1	21.0	14.8	18.5	20.2	28.7	20.2	21.0
			2	19.8	15.1	18.1	18.9	25.9	17.8	19.1
			10	21.0	15.9	18.5	17.4	25.7	13.6	16.2
		5	1	15.4	6.3	11.4	13.4	42.2	14.5	16.2
			2	18.8	7.1	13.9	14.8	41.8	15.5	17.9
			10	20.5	7.5	15.0	12.2	36.0	8.3	11.5
Female rats 19.1%	Leuk./ lymphoma	1	1	55.6	54.6	55.1	55.3	53.6	55.2	55.6
			2	54.3	54.1	53.5	54.2	53.8	53.9	54.0
			10	56.0	55.1	55.7	55.5	54.2	47.5	49.8
		2	1	56.7	48.0	55.4	51.8	73.8	54.5	56.8
			2	53.4	44.1	52.8	46.2	68.5	47.6	48.8
			10	54.1	44.9	52.2	45.4	65.1	26.9	32.9
		5	1	51.2	25.2	49.8	33.8	92.9	45.9	51.4
			2	48.5	24.5	46.5	27.9	91.9	33.9	37.7
			10	49.4	25.0	47.0	25.8	79.8	8.7	14.1

[a] Treatment lethality: $1 + \alpha_0 = 1$ (no treatment lethality); $= 2$ (low lethality); $= 5$ (high treatment lethality).

[b] Tumor lethality: $1 + \phi_0 = 1$ (incidental tumor); $= 2$ (low tumor lethality); $= 10$ (high tumor lethality).

[c] Test statistics: Z_o = Onset test; Z_t = Cochran–Armitage trend test; Z_p = Poly-3 test; Z_{tt} = Truncated trend test; Z_{lt} = Life-table test; Z_{hw} = Hoel–Walburg test; Z_{log} = Logistic regression score test.

[d] Entries are given in percentages.

The Type I error rate of the life-table test, Z_{lt}, increases as the level of treatment lethality increases. Within a given degree of treatment lethality, increasing tumor lethality leads to rejection probabilities that are closer to the nominal levels. This is expected because, as ϕ_0 increases, the age at death from the tumor converges to the age at tumor onset. The empirical behavior of this test matches its predicted behavior based on a consideration of the event-specific hazard functions, $h_i(s)$, given earlier. For high treatment lethality and no tumor lethality, one would expect that the number at risk would decrease with the administered dose. Because of the decreased number at risk in the higher-dose groups, the expected number of tumor-bearing deaths in the higher-dose groups will also be decreased. Hence, the observed minus expected tumor-bearing deaths will increase with dose, which leads to Z_{lt} operating in an anticonservative manner.

The rejection probabilities of the Hoel–Walburg test, Z_{hw}, and the logistic regression test, Z_{log}, were affected in a complex fashion by treatment lethality and tumor lethality. This pattern is expected when one considers equation (3). Within a particular level of tumor lethality, the Type I error rate tended to decrease with increasing levels of treatment lethality as predicted by Lagakos (1982). Dinse (1985) illustrated similar effects of treatment lethality on the prevalence tests Z_{hw} and Z_{log}.

4.2 *Power Results*

Table 2 presents the probability of rejection for the test statistics considered in this study for the situation where the compound induces a doubling in tumor onset in the high-dose group over the control group. For conditions of no treatment lethality and no tumor lethality, all tests have essentially the same power, and all tests show an increase in power as the background tumor rate increases. Within a given level of background tumor rate, the power of these tests will vary according to the Type I error rate given previously. For example, since quantal response tests become conservative with increasing levels of dose-related toxicity, it is expected (and is observed) that these tests are less powerful with the introduction of dose-related toxicity, and therefore these tests have a decreased capability of detecting true differences in tumor incidence between the groups. The inflated Type I error rate for the life-table test leads to a corresponding inflation in power; hence, many compounds may be incorrectly flagged as tumorigenic when this test is used. Tumor lethality, which leads to conservative Type I error rates in the prevalence tests, is translated into reduced power for detecting true tumorigenic differences between the groups.

Essentially all of the tests have a power of 1.0 for tumor sites with a background tumor rate exceeding 19% and a five- or tenfold increase in tumor incidence in the high-dose group relative to the control group. For a 5% background tumor rate, no treatment lethality, and no tumor lethality, the power of the tests is approximately .8 for the fivefold increase in tumor incidence and nearly 1.0 for the tenfold increase in tumor incidence. For a 1% background tumor rate, no treatment lethality, and no tumor lethality, the power of the tests ranges from .4 for the fivefold increase in tumor incidence to .75 for the tenfold increase in tumor incidence. With regard to increases in tumor lethality and increases in treatment lethality, the tests behave in a similar fashion to the results found in Table 1.

5. Discussion

The results from this simulation study indicate the sensitivity of many standard tests for carcinogenicity to treatment lethality and tumor lethality. Thus, treatment lethality and tumor lethality clearly play important roles in the analysis of bioassay experiments.

Quantal response trend tests are robust to tumor lethality assumptions, which is not surprising because these tests depend only on the presence of the tumor and not on the time of tumor occurrence. However, treatment lethality has dramatic effects on the quantal response tests. This is related to the fact that even if a dose response exists, treatment lethality can kill animals prior to the occurrence of a tumor. Since most carcinogenicity studies show some evidence of treatment lethality, a survival-adjusted quantal response trend test is a necessity. The Poly-3 test appears to be superior to the truncated trend test in this regard. The anticonservative nature of Z_p for the low background case described in the previous section results from the fact that the shape of the tumor incidence function for this case is much smaller than 3. A small simulation experiment was done to assess the effect of changes in the shape parameter of the tumor incidence function (η_2) on the Type I error rate of the Poly-3 test. Except for the case of a rare tumor, this test does very well at maintaining the proper Type I error rate. For higher treatment lethality ($\alpha_0 \geqslant 4$) and as the onset distribution shape parameter increased from 3, the Type I error rate dropped off dramatically. We also considered the operating characteristics of an obvious modification of the Poly-3 test, the Poly-k test, when $\eta_2 = k$. We found that, except for extremely small backgrounds, the true Type I error rate was not "significantly" different from the nominal level. Thus, when some knowledge of the shape of the tumor incidence function over time is available, the Poly-3 test can be improved.

The utility of the life-table test is questionable because it is extremely sensitive to treatment lethality. The prevalence tests are proposed as tests that correct for treatment lethality; however, we were surprised to find the degree of the effect that extreme treatment lethality can have on the Type I error rates of these tests. The magnitude of the effect of treatment lethality on the life-table test and on the prevalence tests was seen to depend on tumor lethality. Since many studies will have moderate treatment lethality and unknown tumor lethality, these tests should be used with care.

As with any analysis of the operating characteristics of test statistics, the results of this study are applicable only to the cases considered. However, the cases considered here cover a broad range of possibilities and should be applicable to most carcinogenicity experiments. One case we did not consider was when the effect of treatment on tumor incidence was nonlinear. Since all of the tests studied here assume a linear trend as the alternative hypothesis, we feel justified in considering only the linear case. As further research, it would be of interest to consider the power of the more robust linear trend tests when the data arise from a nonlinear treatment effect. Finally, the logistic regression model used in this analysis controlled for survival differences by using a linear time effect. It may be possible to improve the operating characteristics of this test statistic by using a cubic time effect similar to that used by the Poly-3 test.

We would like to note also that several tests are available for directly controlling the effects of mortality in animal carcinogenicity experiments (McKnight and Crowley, 1984; Dewanji and Kalbfleisch, 1986; Portier, 1986; Portier and Dinse, 1987). All of these procedures require interim sacrifices in addition to a terminal sacrifice. These procedures are likely to have reduced power for testing for increased trends in tumor incidence when compared to the procedures outlined in this paper due to fewer assumptions and an increased number of estimated parameters. However, this decreased power more accurately reflects our lack of knowledge about mortality and tumor incidence. We suggest that researchers consider modifying experimental designs to include interim sacrifices to allow for the use of these newer tests.

In summary, when no information is available on tumor lethality and differences in treatment lethality exist in terminal sacrifice studies, the Poly-3 procedure appears to be the most robust test. If information is available about tumor lethality, a survival-adjusted test can be used. If the shape of the tumor incidence function is expected to follow time to some power k, the Poly-3 test can be modified to become a Poly-k test, which should have superior operating characteristics to the Poly-3 test.

ACKNOWLEDGEMENTS

The authors would like to thank Drs Ken Brown, Gregg Dinse, and Norm Kaplan for their many helpful comments.

RÉSUMÉ

Il est montré que les tests statistiques utilisés en carcinogénèse sont plus ou moins robustes vis à vis des effets de la mortalité. Deux types de mortalité sont considérés: celle induite par la tumeur d'intérêt et celle due au traitement, indépendamment de la tumeur. Des simulations d'éxpériences à faibles effectifs montrent que les deux méthodes le plus couramment utilisées (celle de la table de survie et le test de tendance linéaire de Cochran et Armitage) sont très sensibles à la letalité due au traitement. La letalité tumorale affecte les performances des méthodes habituelles de traitement des pourcentages, telles que la regression logistique. Un test simple sur une réponse en tout ou rien, avec prise en compte de la survie, semble la plus robuste de toutes les procédures considérées.

REFERENCES

Armitage, P. (1955). Tests for linear trend in proportions and frequencies. *Biometrics* **11,** 375–386.

Birch, M. (1965). The detection of partial association, II. The general case. *Journal of the Royal Statistical Society, Series B* **27,** 111–124.

Dewanji, A. and Kalbfleisch, J. (1986). Nonparametric methods for survival/sacrifice experiments. *Biometrics* **42,** 325–342.

Dinse, G. (1985). Testing for a trend in tumor prevalence rates: I. Nonlethal tumors. *Biometrics* **41,** 751–770.

Dinse, G. and Lagakos, S. (1983). Regression analysis of tumor prevalence data. *Applied Statistics* **32,** 236–248.

Gart, J., Chu, K., and Tarone, R. (1979). Statistical issues in interpretation of chronic bioassay tests for carcinogenicity. *Journal of the National Cancer Institute* **62,** 957–974.

Gart, J., Krewski, D., Lee, P., Tarone, R., and Wahrendorf, J. (1987). *Statistical Methods in Cancer Research, Volume* III: *The Design and Analysis of Long-Term Animal Experiments.* Lyon: International Agency for Research on Cancer.

Haseman, J. (1984). Statistical issues in the design, analysis, and interpretation of animal carcinogenicity studies. *Environmental Health Perspectives* **58,** 385–392.

Hitchcock, S. (1966). Tests of hypotheses about the parameters of the logistic function. *Biometrika* **53,** 535–544.

Hoel, D. and Walburg, H. (1972). Statistical analysis of survival experiments. *Journal of the National Cancer Institute* **49,** 361–372.

Lagakos, S. (1982). An evaluation of some two-sample tests used to analyze animal carcinogenicity experiments. *Utilitas Mathematicas* **21B,** 239–260.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50,** 163–170.

McKnight, B. and Crowley, J. (1984). Tests for differences in tumor incidence based on animal carcinogenesis experiments. *Journal of the American Statistical Association* **79,** 639–648.

Peto, R., Pike, M., Day, N., Gray, R., Lee, P., Parish, S., Peto, J., Richard, S., and Wahrendorf, J. (1980). Guidelines for simple, sensitive significance tests for carcinogenic effects in long-term animal experiments. In *Annex to Long-Term and Short-Term Screening Assays for Carcinogens: A Critical Appraisal,* International Agency for Research on Cancer Monographs, **Supplement 3,** 331–426. Lyon: IARC.

Portier, C. (1986). Estimating the tumor onset distribution in animal carcinogenesis experiments. *Biometrika* **73,** 371–378.

Portier, C. and Dinse, G. (1987). Semiparametric analysis of tumor incidence rates in survival/sacrifice experiments. *Biometrics* **43,** 107–114.

Portier, C., Hedges, J., and Hoel D. (1986). Age-specific models of mortality and tumor onset for historical control animals in the National Toxicology Program's carcinogenicity experiments. *Cancer Research* **46,** 4372–4378.

Tarone, R. (1975). Tests for trend in life table analysis. *Biometrika* **62,** 679–682.

Received October 1986; *revised May and November* 1987.

Appendix

Derivation of the Modified Quantal Response Tests

Consider the proportion

$$r_i = \frac{d_{i1}}{n_{i1}^*},$$

where d_{i1} and n_{i1}^* are defined as in Section 2.2 where the data have been collapsed over the entire study period. We know that

$$\lim_{n_{i1}\to\infty} \frac{d_{i1}}{n_{i1}} = \pi_i = \int_0^{TS} \lambda_i(u)F_i(u)\, du,$$

where

$$F_i(u) = \exp\left\{-\int_0^u [\lambda_i(x) + \beta_i(x)]\, dx\right\}.$$

Using the notation from Section 2.2, assume that animals that get the tumor during the course of the study and animals that live to terminal sacrifice are given a weight of $\omega_{ij} = 1$, and let $g_i(t_{ij})$ denote the weight given to animals that die prior to study termination (TS) and are tumor-free, where t_{ij} denotes the age at death for the jth animal in the ith group. It can be shown that

$$\lim_{n_{i1}\to\infty} \frac{n_{i1}^*}{n_{i1}} = \pi_i + F_i(TS) + \int_0^{TS} g_i(u)\beta_i(u)F_i(u)\, du.$$

With this, we are able to calculate what r_i looks like for large samples. After a little algebra, we find that

$$\lim_{n_{i1}\to\infty} (1 - r_i) = S_i(TS)\Omega_i,$$

where

$$S_i(s) = \exp[-\Lambda_i(s)], \quad \Lambda_i(s) = \int_0^s \lambda_i(x)\, dx,$$

$$\Omega_i = \frac{1 - \int_0^{TS} \beta_i(u)F_i(u)[S_i(u)^{-1} - g_i(u)S_i(TS)^{-1}]\, du}{1 - \int_0^{TS} \beta_i(u)F_i(u)[1 - g_i(u)]\, du}.$$

If $\Omega_i = 1$, then r_i approximates $1 - S_i(TS)$, which is a function of only the tumor incidence rate, $\lambda_i(t)$, and is independent of mortality.

It follows that $\Omega_i = 1$ if $S_i(r)^{-1} - g_i(r)S_i(TS)^{-1} = 1 - g_i(r)$ for all r or, equivalently, if

$$g_i(r) = \frac{1 - S_i(r)^{-1}}{1 - S_i(TS)^{-1}}$$

for all r. However, in order to use this weight, one must know the entire dose–time–response model. If we use a first-order Taylor approximation of $S_i(u)^{-1}$, we obtain

$$\hat{g}_i(r) = \frac{\Lambda_i(r)}{\Lambda_i(TS)}.$$

The advantage to using this approximation is that the scale parameter, η_1, of the Weibull hazard described in Section 3 factors out of the formula. The weight is then fully specified by the shape parameter, η_2, and is given by

$$\hat{g}_i(r) = \left(\frac{r}{TS}\right)^{\eta_2}.$$

For some rodent studies, $1 + \Lambda(r)$ will poorly approximate $\exp[\Lambda(r)]$. Therefore, we must be concerned with the degree to which Ω_i differs from 1 when this approximation is used. Substituting

$\hat{g}_i(r)$ for $g_i(r)$ results in $\hat{\Omega}_i = 1 + R_i$, where R_i is given by

$$R_i = \frac{\int_0^{TS} \beta_i(u)F_i(u)\{\sum_{j=2}^{\infty} [\Lambda_i(u)^j - \Lambda_i(u)\Lambda_i(TS)^{j-1}]/j!\}\, du}{1 - \int_0^{TS} \beta_i(u)F_i(u)[1 - \Lambda_i(u)/\Lambda_i(TS)]\, du}\,.$$

By studying the terms within the integral in the numerator, we see that this remainder will be very small in most cases. Since very few animals die early in carcinogenicity experiments, $\beta_i(u)$ is virtually zero for all small values of u. For larger values of u, the difference between $\Lambda_i(u)^j$ and $\Lambda_i(u)\Lambda_i(TS)^{j-1}$ becomes very small. Thus, except for the cases where $\beta_i(t)$ is large for small times, this remainder will be very small. This result is supported by numerical results from both the small-sample simulations and the direct calculation of R_i.

For large samples, a trend test through these r_i values should have the proper operating characteristics for testing the hypothesis of equal tumor incidence rates among the various groups. Although we present no additional analytical support for this contention, the small-sample simulations provide strong evidence that this is the case.

The derivation given here assumes that some knowledge of the form of $\Lambda_i(t)$ is available and is applied correctly. If the wrong functional form for $\Lambda_i(t)$ is chosen, it is possible to create a bias. This is discussed in Section 5 for the class of Weibull hazards we have considered. It is unknown what could happen with other forms of the tumor incidence function.